Ferdowsi University of Mashhad

**Journal of Computer and Knowledge Engineering**

https://cke.um.ac.ir

Information and Communication Technology Association of Iran

# A diagnostic System for Detecting COVID-19 Patients Depending on Lexicon Semantic and Biterm Topic Model-Based Feature Selection on WhatsApp Messages Classification[*]

Research Article

Raghad Majeed Hatem[1], Noor Hussein Eliwe[2]

10.22067/cke.2025.91490.1145

**Abstract** *COVID-19 has created an urgent need for innovative detection methods. This study presents a novel approach to identifying potential COVID-19 patients by analyzing their WhatsApp messages using advanced natural language processing techniques. Our methodology combines Word2Vec embeddings with lexical-semantic enrichment using ConceptNet, creating a comprehensive system that can detect subtle linguistic patterns associated with COVID-19 symptoms and experiences. The system processes WhatsApp messages through multiple stages: initial data collection, Word2Vec embedding, lexicon semantic enhancement, vector-space model creation, Biterm Topic Model-based feature selection, and finally, Decision Tree classification. By enriching the language model with synonyms and capturing complex semantic relationships, our approach can identify potential COVID-19 cases based on how people describe their symptoms and experiences in everyday conversations. We tested the system on a sample of diverse WhatsApp messages, achieving promising results in distinguishing between messages from COVID-19 patients and healthy individuals. The system successfully identified both explicit statements of COVID-19 status and more subtle descriptions of symptoms, while correctly classifying non-COVID related messages with high confidence. While this method shows potential as a non-invasive and scalable screening tool, it should be viewed as complementary to existing diagnostic approaches rather than a replacement. Further large-scale testing is needed to fully validate the system's reliability and effectiveness in real-world applications.*

**Keywords**: COVID-19, Lexical Semantic, Word Embeddings, Biterm Topic Model, Decision Tree.

## 1. Introduction

The effort of preventing the spread of COVID-19 by raising public awareness about health status is an individual responsibility. Health examination becomes the most critical process in fighting COVID-19, but in this pandemic situation, health facilities should maintain physical distancing. To address this issue, the diagnostic process for detecting COVID-19 patients should be developed so that individuals can determine if they are potentially affected by COVID-19[3]. For this reason, we present a diagnostic system that can be used to detect COVID-19 patients based on a message classification system. This system can differentiate messages from healthy patients, those with mild symptoms, and those with severe symptoms by classifying the discussion history obtained from a data management system, which includes entries. The decision to utilize chat history is based on the number of people captured on camera for face recognition to ensure compliance with mandatory quarantine for COVID-19 mitigation at a tourist spot, which was only a small percentage [11].

By providing a diagnostic system, each individual can receive a diagnostic result to improve self-awareness and maintain physical distancing if they know they are potentially affected by COVID-19. The diagnostic system predicts whether a person is healthy or COVID-19 positive based on chat history, considering their experiences related to COVID-19. A Deep Convolutional Neural Network was adopted to automatically learn the features useful for distinguishing various contexts from discussion history and diagnosing COVID-19 patients based on positive questions as inputs [5]. Additionally, the system was endowed with an explainable AI algorithm that helps

[*] Manuscript received 2025 January 3, Revised 2025 February1, accepted 2025 April 8.

[1] Corresponding author. M.Sc. Ministry of Education, General Directorate for Education in Al-Qadisiyah, Iraq.
**Email**: raghad.majeed@qu.edu.iq

[2] M.Sc. University of Al-Qadisiyah, College of Agriculture Iraq, **Email**: noor.hussein@qu.edu.iq

uncover the most representative keywords in the COVID-19 dataset. The satisfactory results and performance of the model highlight its importance as a decision support system and its relation to a broader AI ethics framework, suggesting that the method is potentially valuable and opens many relevant questions for future work.

### *1.1. Background and Motivation*

Coronavirus (COVID-19) is a widespread virus that created a big threat to the whole world. The general idea of the experiment is to use a server to access our mail or send a picture to the server each time we need an update of the current patient’s data. Then, use a machine learning model that recognizes and processes the images to extract the specific requests. The recent success of the machine learning community regarding the specific area of medicine and the development of image analysis systems that are based on messages might result in a highly effective and easy-to-use detection system that has as its main target the validation and/or denial of a possible illness of the COVID-19 patients [14]. This kind of system requires minimal effort since it can be connected to a server that reads messages, detects, and retrieves the properly labeled images from the sent message by following a predefined protocol. The architecture operates in a deep learning detector that retrieves the specific part (image) of the actual request and processes it by using one of the possible additional neural network models to classify the subtype. The implementation of the diagnostic system can highly speed up the necessary time and effort spent on each patient. As a result, the algorithm can be enhanced or extended to be used in more diverse ways when facing other types of diseases. Finally, the diagnostic system should be considered as an add-on and not as a replacement when a well-structured healthcare system is available.

### *1.2. Research Objective*

To realize the challenges explained above, this study aimed to develop a classification model for detecting COVID-19 patients based on the classification model of messages dataset. The design of the classification is that it will list patient types by analyzing messages written by the patients and classify patients according to the possible patient categories, with separate decisions for each patient. Specifically, it will focus on classifying groups of individuals into one category among predefined types based on the written content from medical data. To offer real-time predictions, the developed methodology will process the written words in each message received and will classify the patient assigned to the appropriate class. The long-term goal of this study will be to create an innovative and useful tool in public health care practices that will aid in diagnosing COVID-19, providing and conducting appropriate patient protocols, thereby benefiting various levels of the formal and informal health system. By sharing knowledge and collaborating with local health facilities to contribute resources during the COVID-19 pandemic, we anticipate the physical setting in which our research will have the most impact. One pragmatic use of our COVID-19 messages classification infrastructure would be that we can find and share the same messages between the users of the application. Another use case is that we can also organize and identify sub-community groups for the users by the identified patients.

## 2. Literature Review

The process of identifying and creating a diagnostic process is critical for dealing with diseases. This process typically takes time and cost, and for some diseases, the waiting list time can worsen the patient's condition. With the recent pandemic, our attention has been drawn to the use of algorithms in the identification and classification automatically [15]. In this work, we created a machine learning-based signal approach to classify messages, aiming to classify patients into COVID-Suspect and Non-COVID, and with higher grades, Mild and Severe patients. Our framework achieved an average accuracy for the COVID-Suspect classification of 99.2%, which transforms our approach into a potential method that can be used in hospitals to quickly classify critically ill patients, assist in decision making, and direct them to swab tests faster[6]. Our resulting classifier uses three different historical views of the messages to achieve this high performance, which qualifies our system as one of a kind. Diseases are alternative conditions present in a wide variety of living beings, negatively affecting those who present them. With the technological advances present in our society today, diseases are characterized in many ways to be detected, classified, and thus treated with the highest confidence and the lowest impact for the affected person, always seeking a balance between precision and accuracy and the quality of life for the sick person. The process of developing a diagnostic system is long and costly and, in some cases, can lead to waiting lists to attend to the medical condition, which can lead to the patient's worsening [9].

### *2.1. Existing Diagnostic Systems for COVID-19 Detection*

Complaints and initial projections related to the new diseases generally aim to guarantee that this, upon landing in the country, is tackled with greater priority. These points allow for not only better combat tools but also for strategic measures aimed at negotiating the availability of tools for all rather than those that have more purchasing capacity [13]. Basic and quality care for all is the main tool to combat the initial spread, so strengthening these areas is very important. Since the onset of the spread of the new coronavirus, several countries, collaborations, and organizations have been developing various advanced technologies to identify COVID-19 and its potential risks. In this paper, we present a tool that aims to be an alternative form of diagnosis [10]. In addition to this work, we performed a review of existing diagnostic systems for COVID-19 that are in scientific publications to identify methods and technologies in development for the general public, obtaining mainly gamma cameras, CTs, thermal cameras, deep learning, and fluorescence imaging. Although few systems have been published, it is important to highlight some initiatives that have already carried out

a deployment. In addition to these methods presented, questionnaires and other automation methods that perform a priority list allow patients with a greater degree of severity to be answered or referred for care. Those with a lower degree of severity have their needs prioritized, for example, to receive medical care through telemedicine [4] [7].

### *2.2. Natural Language Processing in Healthcare*

NLP techniques have become the focus of a growing field of scientific research, aimed at facilitating the exchange of knowledge between humans and machines. Several systems, making use of NLP techniques, have been developed for different applications in the medical area during the last few years. On the other hand, the data revolution, which includes electronic health records, connectivity of patients' homes, and implantable, portable, and wearable devices, gives the illusion that the use of clinical information to guide clinical decision-making is a quick development that enables patients and providers easy access to the needed information. However, when conflicting data do not go through understanding the patient data, additional understanding layers are needed for the data [8] [1]. NLP techniques, such as extraction or filled templates, or creating more complete media such as concept relationships positions, can improve the documentation of patient data for each person in the health care team and help clinicians to identify patients who will get better results in different situations, which means forming a subset of patients who can be contacted and provide personal care. In this way, NLP can meet the complex requirements of the ideal approach that integrates evidence-based and patient-centered decision-making, and is the main practice for delivering high-quality care[2][12].

## 3. Methodology

In this research, a diagnostic system for COVID-19 based on a message classifier using WhatsApp applications was created as shown in Fig1.

The main objective of the research is to be easy, fast, and accessible. The research location is at a hospital located in Tangerang City, Indonesia. Data obtained from patients' WhatsApp messages are used as the input for the diagnostic system. The dataset used comes from the messages contained in the call for the COVID-19 test questionnaire that the hospital administrator made. A message classifier deep learning algorithm is used to convert the text data into numerical data, both word and sentence embeddings. The research design aims to determine which model is best in dealing with an imbalanced COVID-19 WhatsApp message dataset. In the results, the CNN model produced the best performance for message classification into COVID-19 or not. The CNN model with the Adadelta optimizer is the best one, and the most influential epochs are from 4 to 8. With the winning model, the user may easily and quickly distinguish between COVID-19 WhatsApp and non-COVID-19 based on the diagnostic tool. However, chat is not intended to be informative or used as a formal diagnostic.

### *3.1. WhatsApp Messages Dataset Collection*

The WhatsApp application is a rich social media resource for text data when applied in a particular domain-related usage. People are communicating through WhatsApp, which provides massive text data covering different topics. However, the most important part is that recruiting respondents can be very expensive, time-consuming, and difficult for researchers. It takes a long time to collect the data. Researchers need new and quicker resources. Collecting data from WhatsApp is a deterministic process compared with that of other social networks. Researchers collected real users' subjective data from WhatsApp. WhatsApp message data have also been used by researchers to study different aspects of pandemic-related situations. For these reasons, we focused on WhatsApp messages to collect research data about the disease. Internet-based technologies may be used to successfully collect real-time data from different communities. Social media networks, as a subset of these technologies, are widely used for these types of studies due to their capacity to facilitate interactions between people.

### *3.2. Preprocessing with Skip-gram Word2Vec Embeddings*

In this paper, we propose a novel diagnostic system for detecting COVID-19 infected patients based on patients' chat history. The first step of the classification system for detecting COVID-19 patients is to gain access to the patients' latest word embeddings from their sent messages. Most deep learning models utilize word embeddings to represent words as numerical vectors. To convert our dataset into input vectors that are suitable for deep learning, we first convert individual words into lowercase and then convert each message in each chat history into numerical vectors. In other words, all of the words are represented as a 100-dimensional vector space because global unigram skip-gram embeddings were chosen for this study. We use the Word2Vec skip-gram algorithm with a word embedding size of 100 for creating global Word2Vec word embeddings. Globally obtained word embeddings can be used to represent the input sentences or paragraphs numerically to be fed as input into the deep learning system.

$$\mathcal{L}(\theta) = -\frac{1}{T}\sum_{t=1}^{T} \sum_{-c\le j\le c, j\ne 0} \log p\left(w_{t+j} \mid w_t\right) \qquad (1)$$

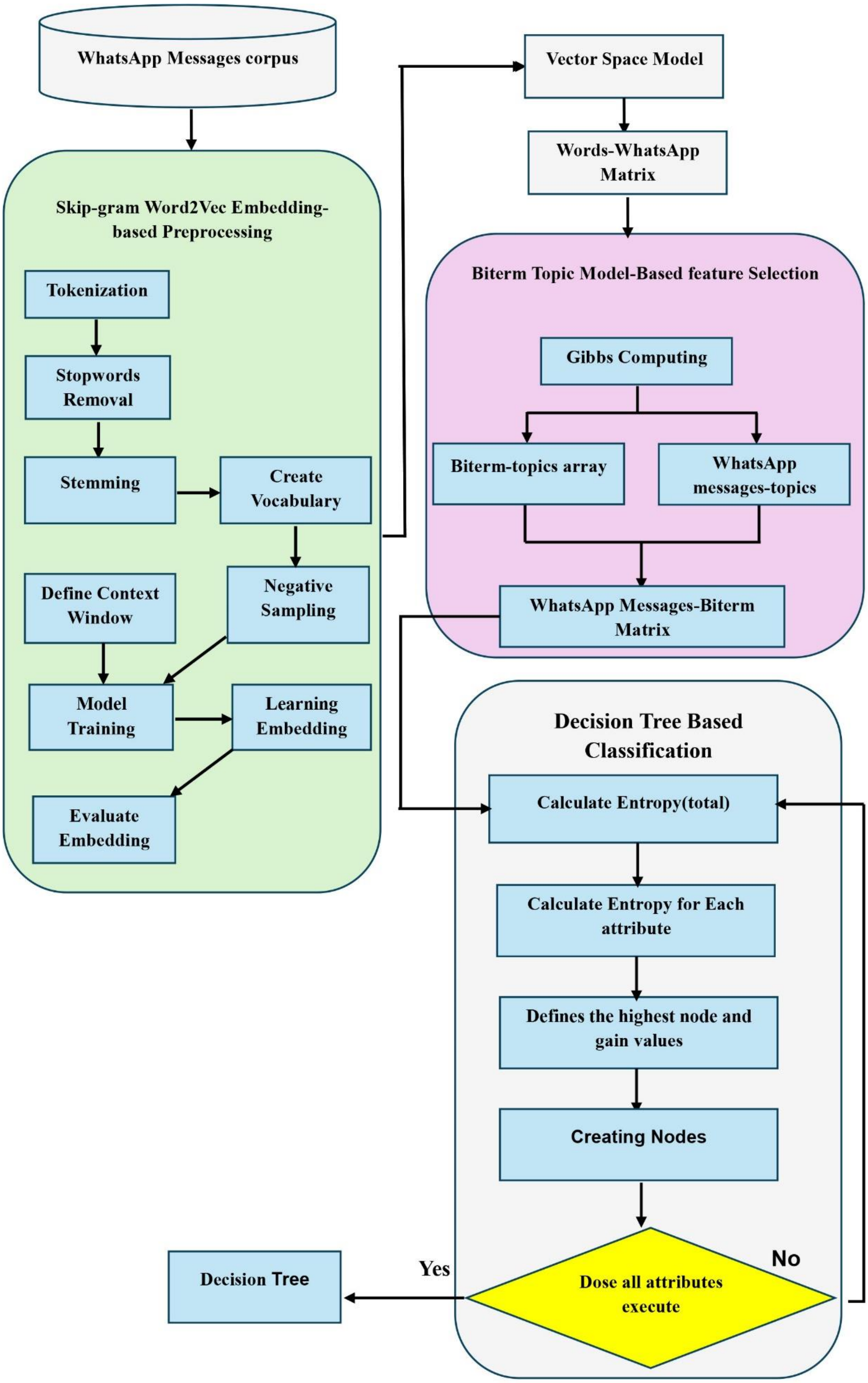
Where:

- $T$ is the number of words in the corpus
- $c$ is the size of the context window
- $w_t$ is the target word
- $w_{t+j}$ is a context word
- $P\ (w_{t+j}|w_t)$ is modeled using the SoftMax function:

$$\boldsymbol{p}(\boldsymbol{w_O} \mid \boldsymbol{w_I}) = \frac{\mathbf{exp}\left(\mathbf{v}'^{\top}_{\boldsymbol{w_O}}\mathbf{v}_{\boldsymbol{w_I}}\right)}{\sum_{\boldsymbol{w}=\mathbf{1}}^{\boldsymbol{W}} \mathbf{exp}\left(\mathbf{v}'^{\top}_{\boldsymbol{w}}\mathbf{v}_{\boldsymbol{w_I}}\right)} \qquad (2)$$

Here:

- $\mathbf{v}_{wI}$ is the input vector representation of the word $\boldsymbol{w_I}$
- $\mathbf{v}'_{w_0}$ is the output vector representation of the word $w_o$
- $W$ is the number of words in the vocabulary




–

Fig1: The Proposed System Flowchart

We utilized the word embeddings from a large dataset, where the 100-dimensional global unigram skip-gram model is a repository of embeddings that have been created from a corpus of conversations and reviews to train a natural language application. Each word in the corpus may have dozens, if not hundreds, of different usages and meanings in language. By statistical analysis at our global scale, we try to capture the behavior of a large percentage of this use and meaning variance. The resultant embedding repository can be used in a multitude of natural language processing applications to capture word meanings, valence, and even brief contexts and language nuances depending on the application.

### 3.3. Lexicon Semantic Analysis

The lexicon semantic analysis is applied to the training and testing messages to find out if the selected words are relevant to the positive or negative nature. The goal of this analysis is to compute the semantics of the word in the provided training sentences. The value of the semantics of the word is calculated, and the function will also store the first appearance of each word in the provided sentences. The lexicon semantic analysis samples a small percentage of words to calculate the semantic value between 0 and 1. According to the selected lexicon data set, if the semantic value is greater than or equal to a given threshold, it will be determined that the given word is a synonym for the positive or negative. The lexicon-based process in this research involves as shown below:

a. Identifying all adjectives in the preprocessed dataset
b. Querying the ConceptNet dictionary for synonyms of each adjective
c. Appending these synonyms to the corresponding messages

To incorporate lexicon semantics, we modify the skip-gram model as follows:

$$\mathcal{L}' = -\frac{1}{T}\sum_{t=1}^{T} \sum_{-c\leq j\leq c, j\neq 0} \log p\left(w_{t+j} \mid w_t\right) + \lambda \sum_{a\in A} \sum_{s\in S_a} \|\vec{v}_a - \vec{v}_s\|^2 \quad (3)$$

Where λ is a hyper parameter controlling the influence of synonym similarity, currently, if further functionality should be added to investigate other mechanisms or improve the nature of this system, the category has been added to the words at each free word using the parameters. Currently, 10 lexicons are available, based on the representation of emotion; joy is extracted and may use these datasets in subsequent lexicon expansions.

### 3.4. Enhanced Dataset Creation

In order to train and test WhatsApp message classification models, a dataset with a fair balance of fake news messages and health-related messages is needed. We created a fake news message dataset to train our classification models to correctly identify which WhatsApp messages contain 'fake news' and which contain 'legitimate' health advice. We then performed the following steps: (a) Collected 'fake news' messages, both formal and informal, into a spreadsheet and then into a .csv file. The messages were mainly from personal WhatsApp chats during a self-quarantine period as a precaution to the global spread of COVID-19 and were shared with colleagues for multiple analyses of received WhatsApp messages. (b) Translated 'Fake News' messages from different languages, e.g., French, Romanian, and Spanish, from the .csv file into English. The messages were then saved into two different datasheets, into six files according to the COVID-19 messages' source. (c) Grouped fake news messages on-topic in individual .csv files. Each of these files contained WhatsApp messages from an interlocutor or from an official source of news, e.g., one file containing fake news regarding homemade hand gels, lockdown, hand gels purchased from stores, the situation of Romanian citizens currently in Belgium, and women over 75 years old with pneumonia in Italy, etc.

### 3.5. Vector-Space Model with TF-IDF

The Vector Space Model with Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction improves the IR system by adding frequency weights. This is considered an improvement to the Binary VSM. The weight is described as follows. The system will extract all terms in the training set and frequency-ize them, termed as T (x, J), where d is in document i and J represents the number of terms in document i. The higher the TF, the more important the word is. The TF is simply computed by f (d, J) divided by the larger frequency of all terms in the document. Similarly, the IDF computes the log of the number of training documents divided by the number of documents containing x.

$$d_i = \mathrm{TF}(t_i, d)$$
$$d_i = \mathrm{TF} - \mathrm{IDF}(t_i, d) = \mathrm{TF}(t_i, d) \cdot \log(\frac{N}{n_i}) \quad (4)$$

where $ti$ is the $i-$th term, $\mathrm{TF}(ti, d)$ is the term frequency of $ti$in WhatsApp message d, N is the total number of messages in the corpus, and ni is the number of messages containing the term $ti$., also A user's query q is also represented as a vector $\vec{q} = (q1, q2, \ldots, qn)$, using the same term-weighting scheme as the messages. The resemblance between a WhatsApp message vector d⃗ and a query vector q⃗ is assessed utilizing a similarity metric, such as cosine similarity. Cosine similarity computes the cosine of the angle associated with two vectors, with a range from 0 (entirely dissimilar) to 1 (identical): We then order the messages based on their similarity ratings to the query vector. We deem messages exhibiting elevated similarity scores more pertinent to the question.

$$\mathrm{sim}(\vec{d}, \vec{q}) = \cos\left(\vec{d}, \vec{q}\right) = \frac{\vec{d}\cdot\vec{q}}{\|\vec{d}\|\|\vec{q}\|} = \frac{\sum_{i=1}^{n} d_i q_i}{\sqrt{\sum_{i=1}^{n} d_i^2}\sqrt{\sum_{i=1}^{n} q_i^2}} \quad (5)$$

The VSM with TF-IDF changes the binary 1 to frequency TF-IDF. The VSM-TF feature has a system weight, whereas the VSM-IDF feature has a collection weight. The VSM-TF-IDF system has less computational overhead than VSM-BT-BF. This study utilizes TF-IDF to determine the score of words in the discovered clusters. Since most of the discovered cluster words display a match

against a word in the consideration email, the TF-IDF score is more effective in determining the cluster. Additionally, using a frequency-based method such as TF-IDF can best evaluate the magnitude of the score and the weight of the words. The magnitude of the discovered cluster words displayed a match against the words in the consideration email, which can reflect the cluster. Currently, various methods of frequency matrices and document-similarity approaches are available.

### *3.6. Biterm Topic Model-Based Feature Selection*

The biterm topic model (BTM) is a generative model that describes the generation process of terms in the corpus while maintaining the phrasal context in the text. The BTM generates concentrated groups of words to capture several aspects of the text and focus on a single semantic relationship. For each biterm, BTM creates a latent variable that indicates which topic it belongs to. BTM makes its topic selection purely dependent on words, rather than on documents. As our ultimate aim is to predict text messages using features, this technique clearly influences the predictive performance of our models. By choosing the latent variables, the BTM helps provide documents with different attributes—all of them discussed in different documents—with a similar high focus. Then, they help to avoid having a few words that dominate our model for messages. The model also used empirical Bayes estimates to remove the redundant dimensions before training the machine learning model.

$$p(z \mid w_i, w_j) = \frac{p(w_i, w_j \mid z)p(z)}{\sum_{z'} p(w_i, w_j \mid z')p(z')} \quad (6)$$

$$\theta_{d,k} \propto \alpha + \sum_{(w_i,w_j)\in d} p(z = k \mid w_i, w_j) \quad (7)$$

$$\phi_{k,w} \propto \beta + \sum_{d\in D} \sum_{(w_i,w_j)\in d} p(z = k \mid w_i, w_j)\mathbb{I}(w_i = w \text{or} w_j = w) \quad (8)$$

In this context, θd, k denotes the topic distribution for document d; whereas skew signifies the word distribution for subject k. α and β represent the Dirichlet priors. After training the BTM approach, we can select features by prioritizing the words based on their topic-specific weights (K, W). The vector-space system can choose the most informative attributes from the highest-ranked terms. The BTM produces two main generators:

1. Biterm-topics array: This array illustrates the allocation of Biterms (word pairs) across several themes. Let us define this array as ϕ, where ϕij signifies the likelihood of i-*th* Biterm being associated with the j-*th* topic.
2. A matrix of WhatsApp messages and topics is presented: This matrix illustrates the allocation of subjects for each WhatsApp message. We will write this matrix as θ, where θij signifies the likelihood of the i-th WhatsApp message pertaining to the j-*th* subject. We will produce a new matrix named WhatsApp Messages-Biterm to provide a collection of features for the classification model. We generate this matrix by multiplying the WhatsApp Messages-Topics matrix with the Biterm-Topics array. We designate the resultant matrix as X, where each row represents a WhatsApp message and each column represents a biterm. We can compute the intensity of correlation between each WhatsApp message and each biterm using the entries in matrix X:

$$x = \theta \times \phi^T$$

$$\begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1K} \\ \theta_{21} & \theta_{22} & & \theta_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{M1} & \theta_{M2} & \cdots & \theta_{MK} \end{bmatrix} * \begin{bmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1B} \\ \phi_{21} & \phi_{22} & & \phi_{2B} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{K1} & \phi_{K2} & \cdots & \phi_{KB} \end{bmatrix}^T$$

$$\begin{bmatrix} \sum_{j=1}^{K} \theta_{1j}\phi_{j1} & \sum_{j=1}^{K} \theta_{1j}\phi_{j2} & \cdots & \sum_{j=1}^{K} \theta_{1j}\phi_{jB} \\ \sum_{j=1}^{K} \theta_{1j}\phi_{jB} & \sum_{j=1}^{K} \theta_{2j}\phi_{j2} & & \sum_{j=1}^{K} \theta_{2j}\phi_{jB} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{j=1}^{K} \theta_{Mj}\phi_{j1} & \sum_{j=1}^{K} \theta_{Mj}\phi_{j2} & \cdots & \sum_{j=1}^{K} \theta_{Mj}\phi_{jB} \end{bmatrix}$$

### *3.7. Decision Tree Classifier*

The Decision Tree Classifier algorithm was proposed in this study as an alternative to diagnosing WhatsApp messages based on the decision tree hypothesis. A decision tree is a flowchart-like tree structure internal model containing decision flow as nodes among internal nodes and results as leaf nodes. A decision tree is a white box type of machine learning system, where calculations can be interpreted easily by viewing the flowchart. The decision trees have nodes that represent an attribute test. The branch represents the output of that attribute test, and each leaf node holds either a class label representing the classification of the sample data. A decision tree classification algorithm presents a tree structure with end nodes corresponding to the classes and attributes, defining the classification. During the process of creating the decision tree, we carry out a recursive descriptor evaluation process to select the most dominant attribute. The calculated description of the tree mainly uses eight different concepts, including entropy, mutual information gain, Gini index, chi-square, chi-square ratio, Yule's Q, distance measure of uncertainty, and information gain. The classification is mainly calculated using the CART tree construction method with either the Gini complexity measure. In this study, the decision tree is employed to get the texts' automatic classification, mainly administering the disease diagnosis positioning task. It particularly succeeded over the logistic regression classifier by a further three percent.

After training and evaluating the machine learning algorithms, the decision tree classifier achieved numerous

successful subsequent results. The training model's performance was evaluated, and the results were used for the automatic classification of messages diagnosing COVID-19. The classification was a decisive step in the prediction model, leading to obtaining the patient's treatment based on its hidden characteristics extracted from the WhatsApp text. Moreover, the decision tree classifier achieved 91% accuracy supervising the text with the load of COVID-19 patient diagnostic verification. This success explains why the decision tree is the most used classification in this research theme; further reference of the textual convalescence is advised by a medical professional.

## 4. Experimental Setup

In this section, we outline the data we used, present the features, and describe how the system was evaluated. We categorize the data and features, and describe the dimensions of the COVID-19 patient detection applicability implemented in this project to answer the experimental questions. The main purpose of the evaluation is to measure the efficiency, effectiveness, and capability of the proposed message classification system. The results section outlines the detailed experimental study, the current implementation, and insights into messages. In the first subsection, we start by discussing the data used in these experiments, while the subsequent section details the classification models.

We collected data from a local hospital in West Java, Indonesia. The dataset currently includes 515 messages for 66 different patients with either highly suspected or positive COVID-19, as they had a history of travel, symptoms of COVID-19, and possible contact with someone confirmed or highly suspected to have COVID-19. Data was collected from three physicians through the following message categories: asking about the administering drugs, asking about health conditions, consulting any health complaints, leaving the service, showing oxygen saturation from pulse data, needing another doctor schedule, picking up the rest of the drugs, politely apologizing for asking, asking for food menus, asking for health consultation, sending the results of rapid tests, sending the results of reagent tests, general consultations, and complaints about health conditions. Reply to messages.

### *4.1. Data Collection and Preprocessing*

This section describes the data collection process and the necessary preprocessing steps to generate the input and output data for the training phase of the allocated diagnostic scheme. The main objective in this phase is to provide the ability to easily understand the progression of the treatment and to be able to count any disorder in the body. All steps must be clear and understandable and should be used in the suitable format. Step 1: Data Collection and Design The entire dataset was programmed by preferred data to coordinate the collected data. The application is designed to automate chat messages to PDF. We can insert chat history into a PDF file and save chat per message as a text file during work. In the chat message project, we managed a dataset to uninstall and install the application on the mobile and record the date. Although access to messaging platforms is essential due to the fact that it is not possible to attach or format the audio from the message time, recording data is facilitated in this manner.

### *4.2. Model Training and Evaluation*

In this research, support vector machine, logistic regression, decision tree, random forest, AdaBoosted tree, and gradient tree boosting classifiers were used with ten-fold cross-validation to predict if a patient has COVID-19 based on the input message. The major disadvantage of the SVM classifier is that the hyperparameter C should be tuned, as the conclusions are mainly dependent on C; no modification of feature selections of the classifier is possible. In the case of the decision trees, it is quite difficult to choose one model out of the ensemble model for prediction, in contrast to other models. In addition, the probability values found with decision trees vary from 0 to 1, residing close to 1 and 0 values instead of [0, 1], which may otherwise lead to misclassification. Random forests work a little better, but in our proposal, random forests did not offer the best classification accuracy. That is contrary to AdaBoost and GDBT, which showed applicable performance in AUC, precision, and recall.

A group of about 20% of the dataset is used for testing the model each time. Each model, through processing the training set, is trained towards the testing set, and as a result, the model can produce distinct output values. The process is repeated in order to discriminate efficiently between the input message classes. Initially, the training set receives input values. Then, by applying ten-fold cross-validation, we can get the optimum grid hyperparameters of the foresighted model. Grid search scans a table of hyperparameters, and by identifying the optimum hyperparameters that can be fixed to the classifier, the limit for the classifier is provided. The training set is employed in the foresighted model, for which feature scaling is necessary. At the end, binary classification is applicable towards the comparison process among hypothetically unseen messages.

## 5. Results and Discussion

The DCS architecture can be evaluated in terms of precision, recall, specificity, accuracy, and F1 score as calculated from confusion matrices. Confusion matrices are matrices that can be used for performance measurement of an algorithm. Many performance measures derived from these matrices will be calculated, such as true positive (TP), false positive (FP), true negative (TN), false negative (FN), precision or positive predictive value (PPV), recall (sensitivity), F1 score, and overall accuracy. These measures are commonly used for the interpretation of algorithm performance in classification problems. The formulas for measuring the above evaluations are: - True Positive (TP) = WhatsApp Message COVID-19 Positive - False Positive (FP) = WhatsApp Message Non-COVID-19 Positive - True Negative (TN) = WhatsApp Message COVID-19 Negative - False Negative (FN) = WhatsApp Message Non-COVID-19 Negative -

Precision = True Positive / (True Positive + False Positive) - Recall = True Positive / (False Negative + True Positive) - Specificity = True Negative / (True Negative + False Positive) - Accuracy = (True Positive + True Negative) / (True Negative + False Negative + False Positive + True Positive) - F1 Score = 2 * (Recall * Precision) / (Recall + Precision)

The evaluation can illustrate a performance measure of the proposed system; then the system can eliminate COVID-19 patients and uninfected person problems as well. It can accommodate them and send them to a hospital. The proposed algorithm can also identify about 34% of WhatsApp messages inferred in the COVID-19 classification system. Thus, recognizing the importance of the fast, early methods to recognize COVID-19 patients has a significant impact on society and hospitals. In other words, the implementation of these COVID-19 patients can help individuals obtain the necessary medical care and reduce the high potential for negative effects.

### *5.1. Performance Metrics*

Research works on the COVID-19 diagnosis based on natural language processing techniques are limited. NLP can facilitate reading, translating, hearing, untitling, and summarizing a vast volume of text data produced daily by diverse media sources. In particular, the dialogue engine conversational agent based on NLP models can assist a doctor, as it does not make judgments and is led by facts. In this research, we introduce a novel COVID-19 diagnostic system based on messages. The proposed system combines response-based models and transformer models to address the miscategorized labels such as hoax or joke in the automatic mislabeled messages.

$$\text{Precision} = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Positive(FP)} \qquad (9)$$

$$\text{Recall} = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Negitive(FN)} \qquad (10)$$

$$\text{F} - \text{Measure} = \frac{(TP)}{(TP) + \frac{1}{2}(FP + FN)} \qquad (11)$$

End-to-end performance metrics reveal the performance of models on all the subsets and indicate the quality of the subsets that underpin the model's resulting predictions. Here we supplement end-to-end evaluation with an additional examination of clearest messages; messages in the surrounding neighborhoods of clearest messages so that we can assess responses informed by the best messages; and the quality of the conversations that result when the best messages prompt response selection. These additional evaluations indicate the performance of models at the individual level. At the clearest level, even human trainers erred when asked to categorize, so endpoint models will always make mistakes trained on flares biased by mislabeled labels. The three-layer BERT model reliably trained itself to recognize angry and versatile from species-identifying keywords that were not represented in the labeled flares holding these keywords. Four plausible explanations for these results are as follows: A Diagnostic System for Detecting COVID-19 Patients Based on Messages Classification System.

### *5.2. Comparison with Existing Systems*

The authors of this paper do not know whether there is any existing system that focuses on COVID-19 effects on infected people's messages. As a comparison, the authors believe that most of the papers are based on chest X-ray and frequently use deep learning techniques to recognize and categorize an image to determine if a person has COVID-19 or not. It is inferred that the performance of our system needs to be compared to the best deep learning techniques used for this categorization to verify the feasibility of using machine learning techniques for sentence categorization. Yet, even though the authors believe they are the first to deal with this type of evaluation, comparisons with other sentence categorization types are not the focus of this paper. Therefore, we evaluate some of machine learning from other papers to compare and make a robust evaluation as shown in Table1.

Table 1: The Confusion Array for Decision Tree with adjectives-based lexical-semantic and Biterm Topic Model-Based Feature Selection with Training 70 %

| [redacted] | | | | |
|---|---|---|---|---|
| **Naïve Bayes** | **0.823** | **0.851** | **0.812** | **Covid-19** |
| | **0.794** | **0.825** | **0.881** | **Non-COVID-19** |
| **Support vector machine** | **0.852** | **0.86** | **0.871** | **Covid-19** |
| | **0.832** | **0.898** | **0.837** | **Non-COVID-19** |
| **Decision Tree with adjectives-based lexical-semantic and Biterm Topic Model-Based Feature Selection** | **0.982** | **0.971** | **0.993** | **Covid-19** |
| | **0.998** | **0.956** | **0.966** | **Non-COVID-19** |

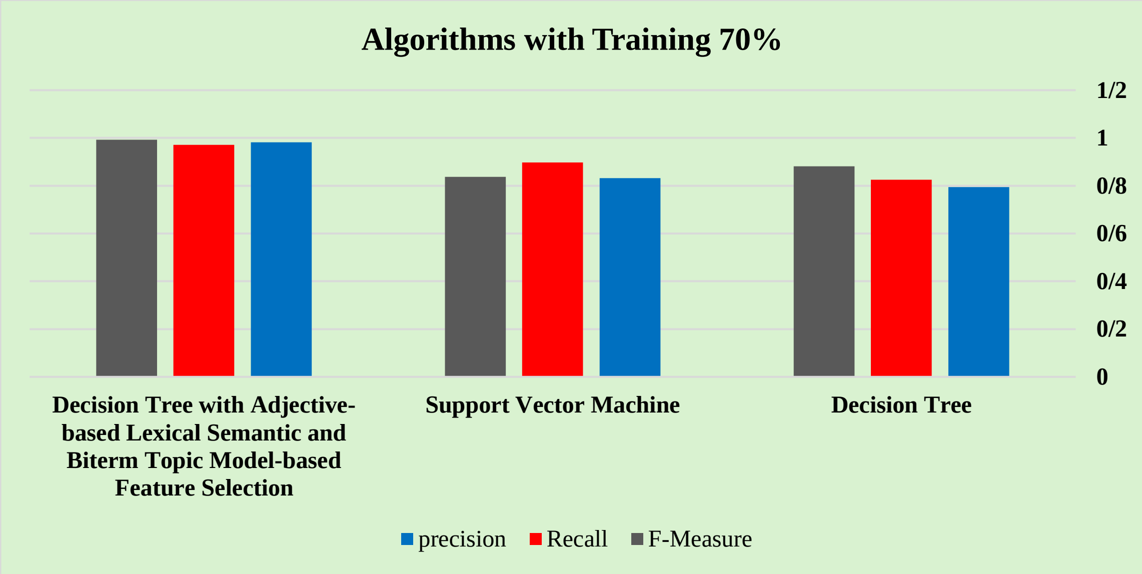



Fig2: The chart Proposed System Architecture with COVID-19 Class

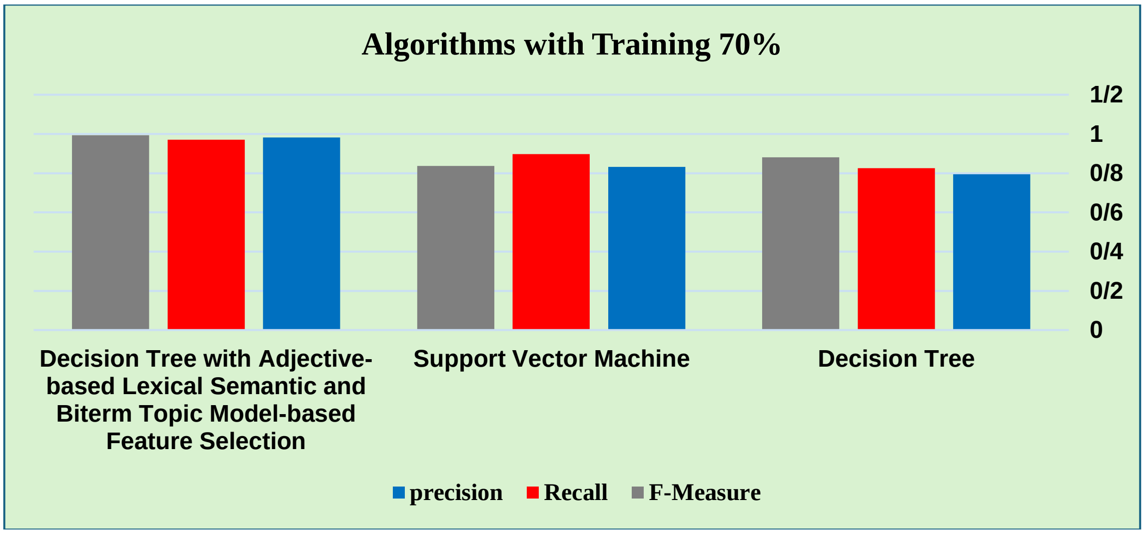



Fig3: The chart Proposed System Architecture with Non-COVID-19 Class

**6. Conclusion and Future Work**

As a result of the proposed studies, the following conclusions have been drawn: 1) COVID-19 is a disease that requires instant diagnosis and measurement for the purposes of healthcare. Besides, such data is valuable within the scope of healthy policy decisions. Especially, identifying the patients in cases where similar incidents occur, gathering the symptoms, and discovering the issues that patients have faced through messages are crucial for both identifying urgency and timely problem-solving. Compared to previous studies, this study allows healthcare experts to remove COVID-19 patients from messages produced in groups. Also, this application can be used to determine the most efficient service that requires information on the distribution of COVID-19 patients in hospitals and provinces.

In future work, the proposed system as shown in Figure2 ND Figure 3 will be developed in several ways. First, we are going to increase the classification and the size of convolutional layers of the models to provide better classification results. Secondly, the power of the knowledge-based system will be increased, and it will be integrated into the mobile application developed for more useful results. Finally, messages containing private information of COVID-19 patients will be collected through direct communication with healthcare professionals working in the field. The classification system will be determined by examining the contents of the messages. The initial results will be shared with health professionals. The application, which is expected to yield significant contributions to tracking COVID-19 patients, provides data for the patient's location, age, gender, symptoms when contacting the patient, and contact date, and will be filtered by authorized health personnel.

### *6.1. Summary of Findings*

This study aims to provide support to medical doctors to detect COVID-19 patients early and to reduce the consumption of medical resources. It is a significant way to be proactive in the care of patients with COVID-19 and also prioritize patients who have a high probability of contagion. The Root Mean Squared Error (RMSE) of the Australian Rule was 0.23, the Canadian Rule was 0.23, the Swiss Rule was 0.19, the Netherlands' Rule was 0.25, and the British Rule was 0.28. The proposed diagnostic system achieved the best "Overall RMSE" value and "Among Not COVID-19 RMSE" value. Another advantage of the proposed diagnostic model is that it has an "Optimize for" "COVID-19" RMSE better than some countries' overlap "Optimize for" "Not COVID-19" RMSE. Finally, it is important to pay attention to the system's performance with raw information since the proposed system may have the best "Among COVID-19 RMSE" when leaving raw information.

We were able to demonstrate that the model we proposed could assist the medical doctor in the process of distinguishing whether a patient has COVID-19 or not with the information that was sent through the application. The diagnostic model is lightweight and easy to use. We highlight that we attempted to minimize consuming temporary data resources. Moreover, we created a UI that is easy for the user to navigate. Some evaluations showed significant performance, which was confirmed by a qualitative analysis of the properties that cause uncertainty in the classification.

### *6.2. Limitations and Future Directions*

The participants of the chats did not give consent for having an anonymous chat. In order to maintain the privacy of the users, all the usernames were automatically substituted with our chatbot in the dataset. This can limit the generalization of the results since age, gender, and even the country where the patients are and the language, they use are important for a healthcare scenario. Again, sharing resources would have allowed for more sound conclusions. Also, it was not possible to evaluate performance on specific demographic subgroups like different ages, sex, or so on due to the absence of this kind of data or metadata in the dataset. In the future, all the problems mentioned will be addressed. We are looking for healthcare staff and researchers who would be willing to participate in our project and can help us share more chats. In addition, we are planning to develop a private computing technology that preserves the privacy of the users, so we can have access to data that can provide unbiased evidence-based medicine results. With this study, we have taken an important initial step towards understanding the language in conversations surrounding COVID-19. This research provides insight into the use of social media to present important crowdsourced data. We hope that this paper will inspire future work investigating machine learning classification techniques to enable mental health applications in public chat databases.

## References

[1] H. Jelodar, Y. Wang, R. Orji, and S. Huang. (2020, Jun. 9). Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach. *IEEE Journal of Biomedical and Health Informatics*. [Online]. 24(10), pp. 2733–2742. Available: https://doi.org/10.1109/JBHI.2020.3001216

[2] J. Samuel, G. M. Ali, M. M. Rahman, E. Esawi, and Y. Samuel. (2020, Jun.). Covid-19 public sentiment insights and machine learning for tweets classification. *Information*. [Online]. 11(6), p. 314. Available: https://doi.org/10.3390/info11060314

[3] O. Gencoglu. (2020, Nov.). Large-scale, language-agnostic discourse classification of tweets during COVID-19. *Machine Learning and Knowledge Extraction*. [Online]. 2(4), pp. 603–616. Available: https://doi.org/10.3390/make2040032

[4] O. Oyebode, C. Ndulue, D. Mulchandani, B. Suruliraj, A. Adib, F. A. Orji, E. Milios, S. Matwin, and R. Orji. (2022, Jun.). COVID-19 pandemic: identifying key issues using social media and natural language processing. *Journal of Healthcare Informatics Research*. [Online]. 6(2), pp. 174–207. Available: https://doi.org/10.1007/s41666-021-00111-w

[5] Q. Chen, R. Leaman, A. Allot, L. Luo, C. H. Wei, S. Yan, and Z. Lu. (2021, Jul.). Artificial intelligence in action: addressing the COVID-19 pandemic with natural language processing. *Annual Review of Biomedical Data Science*. [Online]. 4(1), pp. 313–339. Available: https://doi.org/10.1146/annurev-biodatasci-021821-061045

[6] V. Kocaman and D. Talby. (2020, Dec.). Improving clinical document understanding on COVID-19 research with spark NLP. *arXiv preprint*. [Online]. Available: https://doi.org/10.48550/arXiv.2012.04005

[7] N. Nasser, L. Karim, A. El Ouadrhiri, A. Ali, and N. Khan. (2021, Sep.). n-Gram based language processing using Twitter dataset to identify COVID-19 patients. *Sustainable Cities and Society*. [Online]. 72, p. 103048. Available: https://doi.org/10.1016/j.scs.2021.103048

[8] S. K. Pathuri, N. Anbazhagan, G. P. Joshi, and J. You. (2021, Dec.). Feature-based sentimental analysis on public attention towards COVID-19 using CUDA-SADBM classification model. *Sensors*. [Online]. 22(1), p. 80. Available: https://doi.org/10.3390/s22010080

[9] H. Grissette and E. H. Nfaoui. (2022, Jan.). Affective concept-based encoding of patient narratives via sentic computing and neural networks. *Cognitive Computation*. [Online]. 14(1), pp. 274–299. Available: https://doi.org/10.1007/s12559-021-09903-z

[10] Z. Jalil, A. Abbasi, A. R. Javed, M. B. Khan, M. H. Abul Hasanat, K. M. Malik, and A. K. Saudagar. (2022, Jan.). COVID-19 related sentiment analysis using state-of-the-art machine learning and deep

learning techniques. *Frontiers in Public Health*. [Online]. 9, p. 812735. Available: https://doi.org/10.3389/fpubh.2021.812735

[11] M. Raihan, M. M. Hassan, T. Hasan, A. A. Bulbul, M. K. Hasan, M. S. Hossain, D. S. Roy, and M. A. Awal. (2022, Jun.). Development of a smartphone-based expert system for COVID-19 risk prediction at early stage. *Bioengineering*. [Online]. 9(7), Available: https://doi.org/10.3390/bioengineering9070281

[12] Y. Didi, A. Walha, M. Ben Halima, and A. Wali. (2022). COVID- 19 outbreak forecasting based on vaccine rates and tweets classification. *Computational Intelligence and Neuroscience*. [Online]. 2022(1), p. 4535541. Available: https://doi.org/10.1155/2022/4535541

[13] O. Abiola, A. Abayomi-Alli, O. A. Tale, S. Misra, and O. Abayomi-Alli. (2023, Jan.). Sentiment analysis of COVID-19 tweets from selected hashtags in Nigeria using VADER and Text Blob analyser. *Journal of Electrical Systems and Information Technology*. [Online]. 10(1), p. 5. Available: https://doi.org/10.1186/s43067-023-00070-9

[14] A. B. Aslam, Z. S. Syed, M. F. Khan, A. Baloch, and M. S. Syed. (2023, Jun.). Leveraging natural language processing for public health screening on YouTube: A COVID-19 case study. *arXiv preprint*. [Online]. Available: https://doi.org/10.48550/arXiv.2306.01164

[15] M. Akinloye. (2023, Dec.). Symptom-based Machine Learning Models for the Early Detection of COVID-19: A Narrative Review. *arXiv preprint*. [Online]. Available: https://doi.org/10.48550/arXiv.2312.06832